Statistics - Lecture Three

Charlotte Wickham
wickham@stat.berkeley.edu
http://www.stat.berkeley.edu/~wickham/

Linear Models

1. The Theory
2. Practical Use
3. How to do it in R
4. An example
5. Extensions

Both Rice [2007] and Freedman [2005] give good introductions to the theory of linear models. An excellent resource for applying the theory in R is Faraway [2002]. Venables and Ripley [2002] also offer a lot of tips for using R for linear models and their extensions.

1 The Theory

Consider we preform an experiment and make n measurements of a response variable y. At the same time we also measure a series of predictor variables $x_1, \ldots, x_p$ once for each observation of y. Now we have a array of collected data. We may be interested in how the x can be used to the predict the value of y. In a linear model the parameters enter the model linearly (the predictors do not have to be linear). For example we might believe the model to be,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon,$$

where ϵ is a random variable representing the error and the β_i are unknown parameters. This is an example of a linear model. Another example is,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \epsilon.$$

Note that the predictor is a non-linear function, however this model still falls into the linear model category since the parameters enter linearly. An example of a model that is non-linear would be,

$$Y = \beta_0 + \beta_1 X_1^{\beta_2} + \epsilon.$$

Matrix Representation

It is generally much easier to formulate the models in matrix form. Say we have a linear model of the form

$$y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi} + \epsilon_i.$$

We can write it as

$$Y = X\beta + \epsilon$$

where Y represents a vector of length n containing the observed values $Y = (y_1, ..., y_n)^T$, β is a vector for the parameters $\beta = (\beta_0, \ldots, \beta_p)^T$, ϵ is a vector for the errors $\epsilon = (\epsilon_1, \ldots, \epsilon_p)^T$ and X is a matrix of the predictors,

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1p} \\ 1 & x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{np} \end{pmatrix}.$$

The column of ones in the matrix X incorporates the intercept.

Example - Simple Regression

Say we have observed pairs of values (y_i, x_i) for $i = 1, \ldots, n$. We assume that the relationship between the two variables can be modeled by,

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

where the ϵ_i are iid with mean zero. We can write this in matrix form as

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

Example - Analysis of Variance

In an simple one way analysis of variance we have observed a variable y_{ij} for $i = 1, \ldots, n_j$ which we believe depends on some grouping factor. For example, we might administer treatment j, $j = 1, \ldots, J$ to n_j people and then measure their reaction. To be more concrete, imagine we administer 3 treatments each to five people and observe a measure of their reaction y. We assume the model

$$y_{ij} = \beta_j + \epsilon_{ij}$$

This means that the reaction of person i who receives treatment j depends on the treatment they received plus an individual error (which we will assume is iid with mean zero across treatments). We can write this in matrix form as,

$$\begin{pmatrix} y_{11} \\ y_{21} \\ \vdots \\ y_{51} \\ y_{12} \\ \vdots \\ y_{52} \\ y_{13} \\ \vdots \\ y_{53} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} + \begin{pmatrix} \epsilon_{11} \\ \epsilon_{21} \\ \vdots \\ \epsilon_{51} \\ \epsilon_{12} \\ \vdots \\ \epsilon_{52} \\ \epsilon_{13} \\ \vdots \\ \epsilon_{53} \end{pmatrix}$$

Here the columns of X are acting as dummy variables. The value in the first column of X is 1 if the patient received treatment 1 and 0 otherwise. Similarly with columns 2 and 3. Dummy variables are often used in linear models to introduce an additive effect for a categorical variable. A typical example would be adding a dummy variable for "male" where researchers believe that gender has an additive effect on the variable of interest. In `R` dummy variables are created automatically when factors are in a linear model - more about this later.

Least squares estimation

Geometric Approach

We want to separate the error from the systematic components. One way of looking at this problem is to say we want a solution (our fitted values) that lies in the space spanned by X that is closest to Y. The systematic component $X\hat{\beta}$ is the projection of Y onto the space spanned by X and the residuals are $Y - X\hat{\beta}$. This is illustrated in Figure 1.

Non geometric approach

We might consider a good estimate of β as the one that minimizes the sum of the squared errors $\sum_{i=1}^{n} \epsilon_i^2 = \epsilon^T \epsilon$.

$$\begin{aligned} \sum_{i=1}^{n} \epsilon_i^2 = \epsilon^T \epsilon &= (Y - X\beta)^T (Y - X\beta) \\ &= y^T y - 2\beta X^T y + \beta^T X^T X \beta \end{aligned}$$

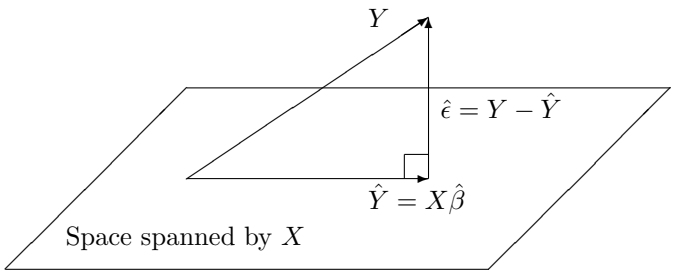


Figure 1: Illustration of the projection approach

Differentiating this and setting to zero we find that the estimate for β that minimizes the squared error satisfies

$$X^T X\hat{\beta} = X^T Y.$$

These are known as the normal equations. The geometric approach gives us the same estimate.

Provided $X^T X$ is invertible,

$$\hat{\beta} = (X^T X)^{-1} X^T Y.$$

We can then get the fitted values, $\hat{Y}$, and residuals, $\hat{\epsilon}$,

$$\hat{Y} = X\hat{\beta} = X(X^T X)^{-1} X^T Y = HY,$$
$$\hat{\epsilon} = Y - X\hat{\beta} = Y - \hat{Y} = (I - H)Y,$$

where the projection matrix $H = X(X^T X)^{-1} X^T$.

Example - Simple regression

Using the notation introduced in the example in the previous section we find,

$$X'X = \begin{pmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{pmatrix} \quad \text{and} \quad X'Y = \begin{pmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i y_i \end{pmatrix}.$$

In order to invert $X'X$ we first need to find its determinant,

$$\begin{aligned} \det X'X &= n\sum_{i=1}^n x_i^2 - \sum_{i=1}^n x_i \sum_{i=1}^n x_i \\ &= n\left(\sum_{i=1}^n x_i^2 - n\bar{x}^2\right) \\ &= n\sum_{i=1}^n (x_i - \bar{x})^2 \end{aligned}$$

Now,

$$\begin{aligned}
\hat{\beta} &= \frac{1}{n\sum_{i=1}^n (x_i-\bar{x})^2}\begin{pmatrix} \sum_{i=1}^n x_i^2 & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & n \end{pmatrix}\begin{pmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i y_i \end{pmatrix} \\
&= \frac{1}{n\sum_{i=1}^n (x_i-\bar{x})^2}\begin{pmatrix} \sum_{i=1}^n x_i^2 \sum_{i=1}^n y_i - \sum_{i=1}^n x_i \sum_{i=1}^n x_i y_i \\ n\sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i \end{pmatrix} \\
&= \frac{1}{n\sum_{i=1}^n (x_i-\bar{x})^2}\begin{pmatrix} \sum_{i=1}^n x_i^2 \sum_{i=1}^n y_i - \sum_{i=1}^n x_i \sum_{i=1}^n x_i y_i + n\bar{x}^2\bar{y} - n\bar{x}^2\bar{y} \\ n\sum_{i=1}^n (x_i-\bar{x})(y_i-\bar{y}) \end{pmatrix} \\
&= \frac{1}{n\sum_{i=1}^n (x_i-\bar{x})^2}\begin{pmatrix} n\sum_{i=1}^n x_i^2\bar{y} - n\bar{x}\sum_{i=1}^n x_i y_i + n\bar{x}^2\bar{y} - n\bar{x}^2\bar{y} \\ n\sum_{i=1}^n (x_i-\bar{x})(y_i-\bar{y}) \end{pmatrix} \\
&= \frac{1}{n\sum_{i=1}^n (x_i-\bar{x})^2}\begin{pmatrix} \bar{y}n\sum_{i=1}^n (x_i-\bar{x})^2 - \bar{x}n\sum_{i=1}^n (x_i-\bar{x})(y_i-\bar{y}) \\ n\sum_{i=1}^n (x_i-\bar{x})(y_i-\bar{y}) \end{pmatrix}
\end{aligned}$$

And from this we get the familiar simple linear regression formulae,

$$\begin{aligned}
\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1\bar{x} \\
\hat{\beta}_1 &= \frac{\sum_{i=1}^n (x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^n (x_i-\bar{x})^2}.
\end{aligned}$$

Exercise - Analysis of variance

Repeat the above example using the set up from the analysis of variance to show that the least squares estimate of β_j is the sample mean of the subjects who received treatment j.

Is $\hat{\beta}$ a good estimate

- Makes sense geometrically
- if the ϵ are normally distributed with constant variance $\hat{\beta}$ is the maximum likelihood estimate
- The Gauss Markov Theorem says it is the best linear unbiased estimator (BLUE)

Mean and Variance of $\hat{\beta}$

$$\begin{aligned}
\mathrm{E}(\hat{\beta}) &= E((X^TX)^{-1}X^TY) = (X^TX)^{-1}X^TX\beta = \beta \\
\mathrm{Var}(\hat{\beta}) &= E\left[(X^TX)^{-1}X^TY\left((X^TX)^{-1}X^TY\right)^T\right] \\
&= \mathrm{E}\left[(X^TX)^{-1}X^TYY^TX(X^TX)^{-1}X^TY)\right] \\
&= \left[(X^TX)^{-1}X^T\sigma^2IX(X^TX)^{-1}X^TY)\right] \\
&= \sigma^2(X^TX)^{-1}
\end{aligned}$$

Estimable

$\phi = c^T\beta$ is estimable if there exists a linear combination a^TY such that

$$\mathrm{E}(a^TY) = c^T\beta$$

If X is full rank then all linear combinations are estimable

Gauss Markov Theorem

Assume the usual set up that $Y = X\beta + \epsilon$, $\mathrm{E}\epsilon = 0$ and $\mathrm{Var}\epsilon = \sigma^2 I$. Let $\phi = c^T\beta$ be an estimable function. Then in the class of all linear unbiased estimators of ϕ, $\hat{\phi} = c^T\hat{\beta}$ has the minimum variance and is unique.

Proof

From David Freedman's Statistical Models.

Let $\tilde{\phi}$ be a competing linear unbiased estimator of ϕ. Since $\tilde{\phi}$ is linear we can write it as $\tilde{\phi} = d^T Y$. We also note that since $\tilde{\phi}$ is unbiased,

$$\mathrm{E}(d^T Y) = d^T \mathrm{E}Y = d^T X\beta = c^T\beta \quad \implies d^T X = c^T.$$

Now we define $q = d - X(X^T X)^{-1}c$ so,

$$q^T = d^T - c^T(X^T X)^{-1}X^T.$$

Multiplying both sides on the right by X

$$\begin{aligned} q^T X &= d^T X - c^T(X^T X)^{-1}X^T X \\ &= c^T - c^T = 0_{1\times p} \end{aligned}$$

Then,

$$\begin{aligned} \mathrm{Var}(\tilde{\phi}) &= \mathrm{Var}(d^T Y) \\ &= \mathrm{Var}(d^T\epsilon) \\ &= \sigma^2 d^T d \\ &= \sigma^2(q^T + c^T(X^T X)^{-1}X^T)(q + X(X^T X)^{-1}c^T) \\ &= \sigma^2(q^T q + c^T(X^T X)^{-1}c) \quad \text{cross products drop since } q^T X = 0 \\ &= \sigma^2 q^T q + var(\hat{\phi}). \end{aligned}$$

Since $q^T q \geq 0$ we have shown that $\hat{\phi}$ has the smaller variance with equality only if $\tilde{\phi} = \hat{\phi}$.

Estimating σ

Can show $\mathrm{E}(\hat{\epsilon}^T\hat{\epsilon}) = \sigma^2(n-p)$ (see for example Freedman, 2005 pg 48), so an unbiased estimate of σ^2 is

$$\hat{\sigma^2} = \frac{\hat{\epsilon}^T\hat{\epsilon}}{n-p}.$$

Distributional Assumptions

We now add the additional assumption that $\epsilon \sim \mathcal{N}(0, \sigma^2)$. This implies directly that,

$$\begin{aligned} Y &\sim \mathcal{N}(X\beta, \sigma^2), \\ \hat{\beta} &\sim \mathcal{N}(\beta, \sigma^2(X^T X)^{-1}). \end{aligned}$$

We can also prove the following results:

$$\hat{Y} \sim \mathcal{N}(X\beta, \sigma^2 H)$$
$$\hat{\epsilon} \sim \mathcal{N}(0, \sigma^2(I - H))$$
$$\hat{\sigma^2} \sim \frac{\sigma^2}{n-p}\chi^2_{n-p}$$

Inference on β

Individual parameters

Under the Normal assumption

$$\frac{\hat{\beta}_i - \beta_i}{s_{\beta_i}} \sim t_{n-p}$$

where s_{β_i} is the standard error of $\hat{\beta}$ and is the square root of the ith diagonal entry of $\sigma(X^T X)^{-1}$. This allows us to preform hypothesis tests and create confidence intervals for the parameters individually. For, example it is common to test the hypothesis that $\beta_i = \beta_{i0}$. When $\beta_{i0} = 0$ rejecting this hypothesis implies that the predictor associated with β_i has a linear relationship with the response. The test statistic for this case is,

$$t = \frac{\hat{\beta} - \beta_{i0}}{s_{\beta_i}}.$$

Under the null hypothesis t follows the Students-t distribution with $n - p$ degrees of freedom. So, for a test at level α we reject if $|t|$ is above $t_{n-p}(\alpha/2)$. Note that due to the duality of confidence intervals and hypothesis tests this is equivalent to the $100(1-\alpha)\%$ confidence interval

$$\hat{\beta} \pm t(\alpha/2)_{n-p} s_{\beta_i}$$

containing β_{i0}.

Parameters jointly

We also might be interested in testing whether a number of parameters are jointly zero. Imagine we have p parameters and want to test if the last p_0 parameters are zero. To do this we use the F test. This involves fitting two models: one in with all parameters, and one with the last p_0 parameters constrained to zero. Let $\hat{\beta}$ be the least squares estimate from the full model and $\hat{\beta}^*$ be the least squares estimate from the smaller model. The the F-statistic is,

$$F = \frac{(||X\hat{\beta}||^2 - ||X\hat{\beta}^*||^2)/p_0}{||\hat{\epsilon}||^2/(n-p)}$$

where $\hat{\epsilon}$ are the residuals from the full model. Equivalently (check this yourself) we can write this as

$$F = \frac{(||\hat{\epsilon}||^2 - ||\hat{\epsilon}^*||^2)/p_0}{||\hat{\epsilon}||^2/(n-p)},$$

where $\hat{\epsilon}^*$ are the residuals from the smaller model. Often $||\hat{\epsilon}||^2$ is referred to as the residual sum of squares (RSS). Under the null hypothesis that the last p_0 parameters are zero this is distributed as $F_{p_0, n-p}$. Often as part of the output of standard statistical programs an F-test will be preformed under the null that all the parameters are zero except for an intercept. A significant result does not validate that the model is correct it just says that it is a very unusual result under the assumption all the parameters are zero.

2 In Practice

- Plot the data
  - Is there anything unusual?
  - Are there signs of collinearity?
  - What models might be appropriate?
- Decide on some appropriate models and hypotheses to test
  - What questions do you want to answer?
  - Does theory give you any hints?
- Fit the models
- Check the fit of the models - adjust if necessary
  - Check assumptions
  - Plot residuals against fitted values
  - Plot residuals against included variables
  - Plot residuals against omitted variables
  - Plot residuals against time
  - Plot qqplot of residuals
  - Plot data and fitted model
- Test hypotheses
- Interpret results - in original terms!
- **Remember association does not imply causality**

3 How to do it in R

In order to give some setting to the following discussion imagine we have observed three variables x, y and z. We would like to use x and z to predict y. I have in fact made up some data: Scatter plots for each pair of variables are plotted by `pairs`. This is something you probably want to do as soon as you have your data.

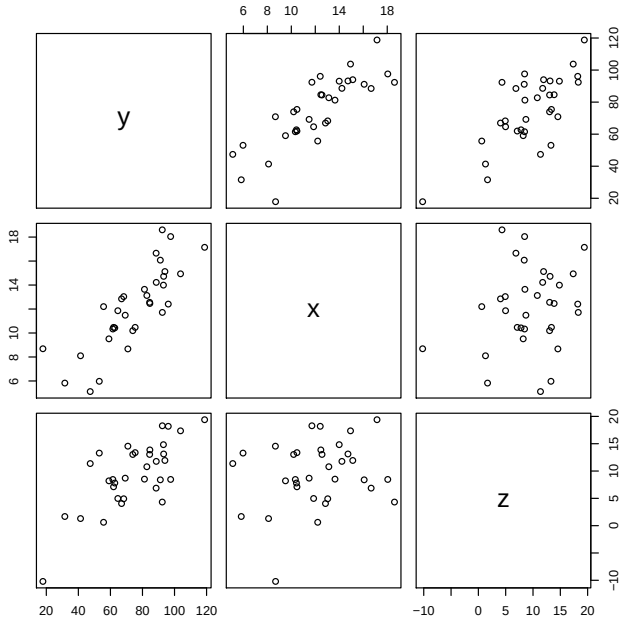
lm

The basic workhorse of fitting linear models in `R` is the function `lm`. The function takes a formula in terms of the variable names. In its simplest form the formula is response ~ predictor. The tilde separates the response variables from the predictors. The simplest example of its use is `lm(y ~ x )` . If we have vectors `x` and `y` the result of the call to `lm` will be the solution to the simple linear regression of x on y.

Examples

`lm` takes a number of other arguments that will specify a dataframe that the data is in, a subset of the data to use, how to deal with missing values and weights (for weighted least squares). Check the help for more details.

```
> x <- rnorm(33, mean = 12, sd = 4)
> z <- rnorm(33, mean = 8, sd = 6)
> y <- 4.5 * x + 2.1 * z + rnorm(33)
> fake.data <- data.frame(y, x, z)
> pairs(fake.data)
```



R Notation	Model being fit
`lm(y ~ x)`	$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$
`lm(y ~ x + z)`	$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \epsilon_i$
`lm(y ~ x - 1)`	$y_i = \beta_1 x_i + \epsilon_i$
`lm(y ~ x:z)`	$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \beta_3 x_i z_i + \epsilon_i$
`lm( log(y) ~ x + I(x^2))`	$\log(y_i) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i$

lm objects

You can save a fitted object by simply assigning it to a variable.

```
> fit <- lm(y ~ x + z, data = fake.data)
```

`R` provides lots of useful functions that you can apply to the fitted object. For example, `summary` provides a standard statistical summary of the fit.

```
> summary(fit)

Call:
lm(formula = y ~ x + z, data = fake.data)

Residuals:
```

```
    Min      1Q  Median      3Q     Max
-2.7650 -0.8891 -0.1078  0.6273  2.3831

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.66049    0.91064  -0.725    0.474
x            4.54637    0.06836  66.507   <2e-16 ***
z            2.11552    0.04177  50.651   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.264 on 30 degrees of freedom
Multiple R-Squared: 0.9952,        Adjusted R-squared: 0.9949
F-statistic:  3100 on 2 and 30 DF,  p-value: < 2.2e-16
```

We can change our model by using the function `update`. The first argument supplies the fitted model we wish to change, the second the new formula. A `.` represents everything that was already in the formula on that side of the ~.

```
> fit2 <- update(fit, . ~ . - 1)
> fit2$call

lm(formula = y ~ x + z - 1, data = fake.data)

> fit3 <- update(fit2, . ~ . - z)
> fit3$call

lm(formula = y ~ x - 1, data = fake.data)

> fit4 <- update(fit, log(.) ~ .)
> fit4$call

lm(formula = log(y) ~ x + z, data = fake.data)
```

The functions `coefficients`, `fitted.values` and `residuals` return the obvious. `predict` also does the obvious. Give it a fitted model and and a set of new data points and it will return predicted values.

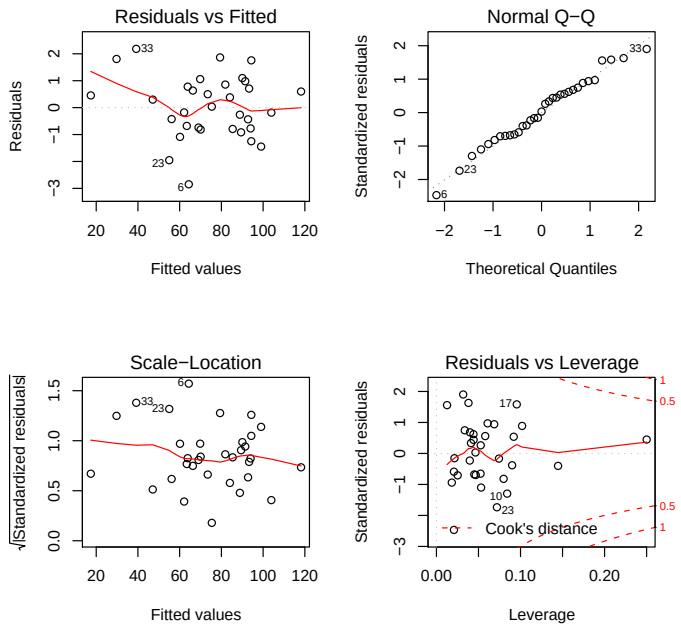
The function `anova` provides two useful functions. If you give it one fitted model it will lay out the results of the fit in an anova table. If you give it two fitted models it will preform an F-test comparing them and put the results in an anova table.

```
> anova(fit)

Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)
x          1 5809.1  5809.1  3635.0 < 2.2e-16 ***
z          1 4099.9  4099.9  2565.5 < 2.2e-16 ***
Residuals 30   47.9     1.6
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> anova(fit3, fit2)

Analysis of Variance Table

Model 1: y ~ x - 1
Model 2: y ~ x + z - 1
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     32 4813.2
2     31   48.8  1    4764.4 3027.5 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Simply using `plot` on a model object will plot a variety of diagnostics

```
> par(mfrow = c(2, 2))
> plot(fit2)
```

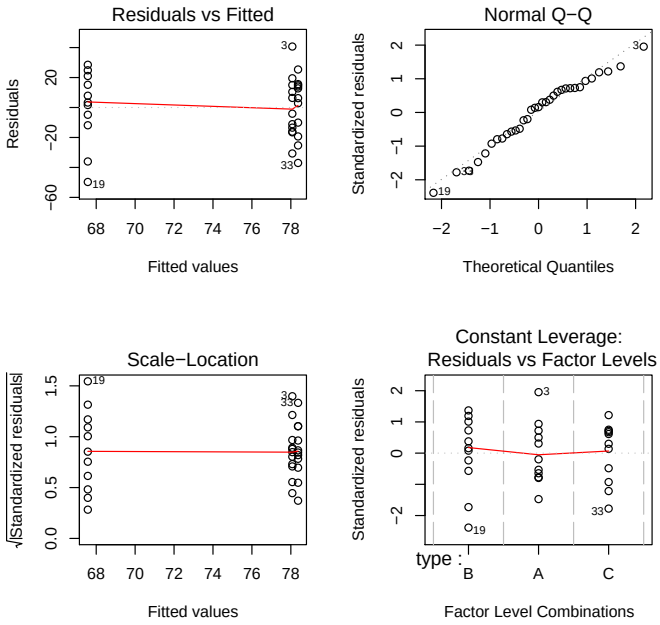


Factors

Factors are Rs way of encoding categorical variables. If a factor is given in the predictor side of the formula, R will automatically create dummy variables that correspond to each level of the factor and estimate the values for the fixed effects.

```
> type <- rep(c("A", "B", "C"), rep(11, 3))
> type

 [1] "A" "A" "A" "A" "A" "A" "A" "A" "A" "A" "A" "B" "B" "B" "B" "B" "B" "B" "B"
[20] "B" "B" "B" "C" "C" "C" "C" "C" "C" "C" "C" "C" "C" "C" "C"
```

```
> str(type)

 chr [1:33] "A" "A" "A" "A" "A" "A" "A" "A" "A" "A" "A" "B" "B" "B" "B" "B" ...

> type <- as.factor(type)
> type

 [1] A A A A A A A A A A B B B B B B B B B B C C C C C C C C C C C
Levels: A B C

> str(type)

 Factor w/ 3 levels "A","B","C": 1 1 1 1 1 1 1 1 1 1 ...

> fit5 <- lm(y ~ type)
> summary(fit5)

Call:
lm(formula = y ~ type)

Residuals:
     Min       1Q   Median       3Q      Max
-31.5729 -13.0392  -0.7073  11.2880  42.0192

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   63.679      5.071  12.558 1.78e-13 ***
typeB         11.439      7.171   1.595    0.121
typeC         -4.409      7.171  -0.615    0.543
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 16.82 on 30 degrees of freedom
Multiple R-Squared: 0.1478,        Adjusted R-squared: 0.09101
F-statistic: 2.602 on 2 and 30 DF,  p-value: 0.09077

> anova(fit5)

Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value  Pr(>F)
type       2 1471.9   735.9  2.6019 0.09077 .
Residuals 30 8485.2   282.8
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> par(mfrow = c(2, 2))
> plot(fit5)
```



4 Example - Analysis of Covariance

Taken from Venables and Ripley [2002].

Preliminary plot

Mr Derek Whiteside of the UK Building Research Station recorded the weekly gas consumption (*Gas* in 1000s of cubic feet) and average external temperature (*Temp* in Celsius) at his own house in south-east England for two heating seasons, one of 26 weeks before, and one of 30 weeks after cavity-wall insulation was installed. The object of the exercise was to assess the effect of the insulation on gas consumption.

Figure 2 shows a scatter plot of the data. We clearly see that the cavity wall insulation appears to have reduced the gas consumption but it also appears to affect the slope of the relationship between the gas consumption and temperature. We would like to formalize this with a model.

Model Proposal

Firstly define *Before* to be a dummy variable indicating a measurement before the insulation was installed and *After* similarly. A linear model seems appropriate so there are two possibilities for modeling the data. The first,

$$\text{Gas}_i = \beta_0 \text{Before}_i + \beta_1 \text{After}_i + \beta_2 \text{Temp}_i + \epsilon_i$$

assumes the slope of the relationship between the gas consumption and temperature is the same before and after insulation is installed but that they have different intercepts. The second,

$$\text{Gas}_i = \gamma_0 \text{Before}_i + \gamma_1 \text{After}_i + \gamma_2 \text{Before}_i \times \text{Temp}_i + \gamma_3 \text{After}_i \times \text{Temp}_i + \epsilon_i$$

```
> summary(whiteside)
```

```
    Insul         Temp             Gas
 Before:26   Min.   :-0.800   Min.   :1.300
 After :30   1st Qu.: 3.050   1st Qu.:3.500
             Median : 4.900   Median :3.950
             Mean   : 4.875   Mean   :4.071
             3rd Qu.: 7.125   3rd Qu.:4.625
             Max.   :10.200   Max.   :7.200
```

```
> print(qplot(Temp, Gas, glyph = Insul, data = whiteside))
```

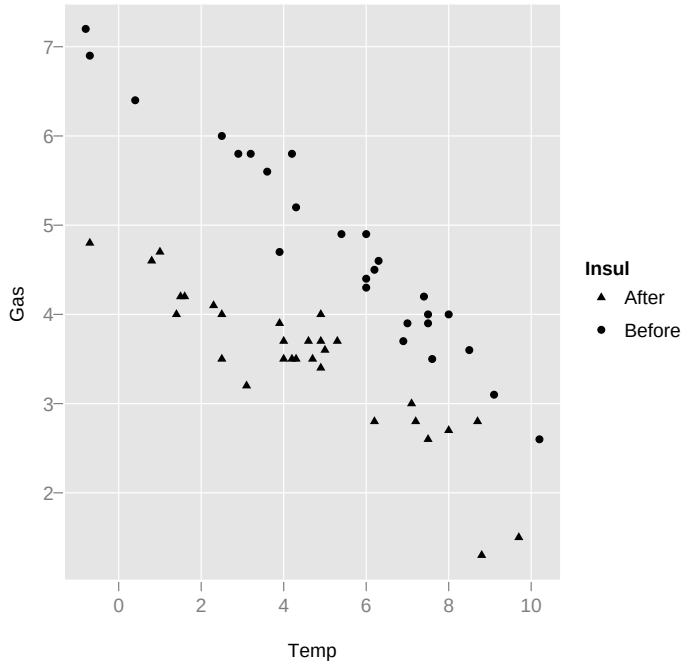


Figure 2: Scatter plot of Temperature versus Gas consumption before and after the installation of cavity wall insulation.

has both different slopes and intercepts for before and after the insulation was installed. We will be able to use an F-test to help choose between these two models.

Model Fitting

```
> fit1 <- lm(Gas ~ Insul + Temp - 1, data = whiteside)
> fit2 <- lm(Gas ~ Insul/Temp - 1, data = whiteside)
> summary(fit1)

Call:
lm(formula = Gas ~ Insul + Temp - 1, data = whiteside)

Residuals:
     Min       1Q   Median       3Q      Max
-0.74236 -0.22291  0.04338  0.24377  0.74314

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
InsulBefore  6.55133    0.11809   55.48   <2e-16 ***
InsulAfter   4.98612    0.10268   48.56   <2e-16 ***
Temp        -0.33670    0.01776  -18.95   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3574 on 53 degrees of freedom
Multiple R-Squared: 0.9933,        Adjusted R-squared: 0.9929
F-statistic:  2600 on 3 and 53 DF,  p-value: < 2.2e-16

> summary(fit2)

Call:
lm(formula = Gas ~ Insul/Temp - 1, data = whiteside)

Residuals:
     Min       1Q   Median       3Q      Max
-0.97802 -0.18011  0.03757  0.20930  0.63803

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
InsulBefore      6.85383    0.13596   50.41   <2e-16 ***
InsulAfter       4.72385    0.11810   40.00   <2e-16 ***
InsulBefore:Temp -0.39324    0.02249  -17.49   <2e-16 ***
InsulAfter:Temp  -0.27793    0.02292  -12.12   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.323 on 52 degrees of freedom
Multiple R-Squared: 0.9946,        Adjusted R-squared: 0.9942
F-statistic:  2391 on 4 and 52 DF,  p-value: < 2.2e-16

> anova(fit1, fit2)
```

```
Analysis of Variance Table

Model 1: Gas ~ Insul + Temp - 1
Model 2: Gas ~ Insul/Temp - 1
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     53 6.7704
2     52 5.4252  1    1.3451 12.893 0.0007307 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The null hypothesis of the F-test is that $\gamma_2 = \gamma_3 = \beta_2$. The p-value from the test is very small giving us strong evidence against this null hypothesis. We therefore conclude that the model with separate slopes is most appropriate. The result from this fit is shown in Figure 3. The fit of the

```
> p <- qplot(Temp, fitted.values(fit2), id = Insul, data = whiteside,
+     type = "line")
> p$ylabel = "Gas"
> print(ggpoint(p, aes = list(x = Temp, y = Gas, shape = Insul)))
```

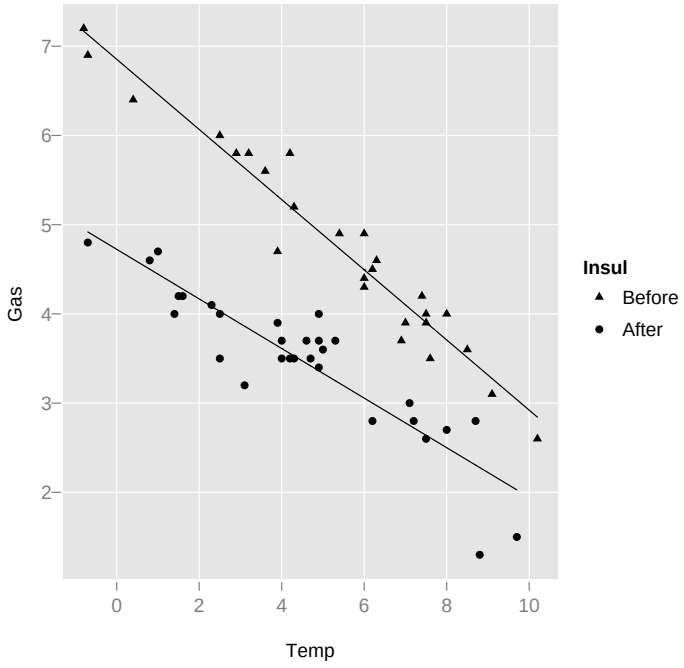


Figure 3: Data with fitted line overlaid

lines looks good. We might suspect a quadratic fit would be better. Individual quadratic lines for before and after installation could be fitted with `fit3<-lm(Gas ~ Insul/(Temp+I(Temp^2)) - 1, data=whiteside)`. This was found to be not significantly better than the linear model. Figure 4 shows diagnostic plots for this model. There is nothing of concern.

```
> par(mfrow = c(2, 2))
> plot(fit2)
```

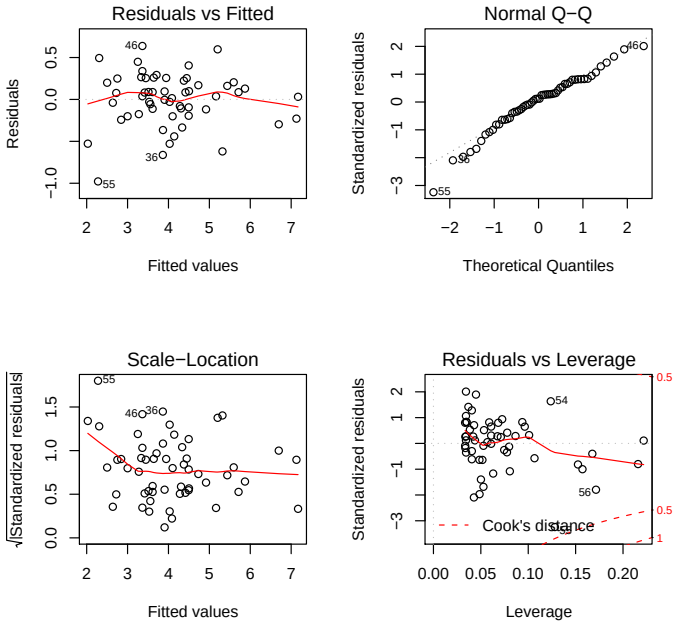


Figure 4: Diagnostics for fitted model.

Interpretation

An equivalent way of specifying our model would have been `fit4<-lm(Gas ~ Insul*Temp, data=whiteside)`. In this model we now have coefficients of the slope and intercept before the insulation and coefficients for the change in slope and intercept after the insulation. This in fact is more interpretable. We can use the information from the fit to calculate 95% confidence intervals for the estimated coefficients.

```
> fit4 <- lm(Gas ~ Insul * Temp, data = whiteside)
> summary(fit4)

Call:
lm(formula = Gas ~ Insul * Temp, data = whiteside)

Residuals:
     Min       1Q   Median       3Q      Max
-0.97802 -0.18011  0.03757  0.20930  0.63803

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      6.85383    0.13596  50.409  < 2e-16 ***
InsulAfter      -2.12998    0.18009 -11.827 2.32e-16 ***
Temp            -0.39324    0.02249 -17.487  < 2e-16 ***
InsulAfter:Temp  0.11530    0.03211   3.591  0.00073 ***
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.323 on 52 degrees of freedom
Multiple R-Squared: 0.9277,        Adjusted R-squared: 0.9235
F-statistic: 222.3 on 3 and 52 DF,  p-value: < 2.2e-16

> est <- summary(fit4)$coefficients[, 1]
> se <- summary(fit4)$coefficients[, 2]
> est

    (Intercept)      InsulAfter            Temp InsulAfter:Temp
      6.8538277      -2.1299780      -0.3932388       0.1153039

> se

    (Intercept)      InsulAfter            Temp InsulAfter:Temp
     0.13596397      0.18009172      0.02248703      0.03211212

> est + qt(0.975, df = 52) * se

    (Intercept)      InsulAfter            Temp InsulAfter:Temp
      7.1266594      -1.7685976      -0.3481153       0.1797416

> est - qt(0.975, df = 52) * se

    (Intercept)      InsulAfter            Temp InsulAfter:Temp
     6.58099603     -2.49135850     -0.43836236      0.05086618
```

Before installing the insulation we estimate with 95% confidence that at zero degrees Celsius the gas consumption is between 6.58 and 7.13 thousand cubic feet. After installation the gas consumption at zero degrees is estimated to decrease by 1.77 to 2.49 thousand cubic feet.

Before installing the insulation we estimate with 95% confidence that for every increase of one degree celsius the gas consumption decreases by between 0.35 and 0.44 thousand cubic feet. After installation this decrease is between 0.05 and 0.18 thousand cubic feet less.

We conclude that over the temperatures observed the insulation did indeed reduce gas consumption.

5 Extensions

The classical linear model makes three assumptions:

First Moment assumption $\mathrm{E}Y = X\beta$

Second Moment assumption $\mathrm{Var}\epsilon = \sigma^2 I$

Distributional Assumption $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$.

Extensions to the classical linear model involve relaxing one or more of these assumptions. Below I discuss some of these extensions but do not give the details on fitting these models (all are discussed in 215B). Venables and Ripley [2002] cover the implementation of all these techniques and provide references if you are interested in the theory.

Generalized least squares

Generalized least squares allows the relaxation of the second moment assumption. It becomes,

$$\text{Var}\epsilon = \sigma^2 V,$$

where V is some known variance covariance matrix. So, now we allow the errors to be correlated and possibly have unequal variances. If V is unknown feasible generalized least squares can be used to estimate its value.

Non linear least squares

Non linear least squares relaxes the first moment assumption,

$$\text{E}Y = \mu(X, \beta),$$

where μ is some known function dependent on unknown parameters β.

General linear models

General linear models (glm) extend linear models to allow the response variables to be non normally distributed. We define the linear predictor to be $\nu = \beta_1 x_1 + \ldots + \beta_p x_p$. The assumptions of the glm are:

- The predictor variables may only influence the distribution of y through ν.
- The distribution of the response, y, needs to belong to the exponential family.
- E(Y) is a smooth invertible function of the linear predictor.

On of the big advantages of this method is that is allows us to deal with categorical response variables. Logit and probit are examples of general linear models.

Generalized additive models

Generalized additive models are very general technique. We assume that,

$$\text{E}Y = \sum_{j=1}^{p} \mu_j(X_j),$$

where the μ_j are unknown smooth functions.

References

Julian J. Faraway. *Practical Regression and Anova using R*. Pdf format, 2002. URL `http://www.stat.lsa.umich.edu/~faraway/book/`.

David A. Freedman. *Statistical Models: Theory and Practice*. Cambridge University Press, 2005.

John A. Rice. *Mathematical Statistics and Data Analysis*. Duxbury Press, Belmont, CA, third edition, 2007.

W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer, New York, fourth edition, 2002.